Three Essays on Ethics : I UTILITY AND PREFERENCES/ II DARWIN ON THE EVOLUTION OF MORALITY/ III SIDGWICK'S THREE PRINCIPLES AND HARE'S UNIVERSALIZABILITY

AUTHOR(S):

Uchii, Soshichi

CITATION:

Uchii, Soshichi. Three Essays on Ethics : I UTILITY AND PREFERENCES/ II DARWIN ON THE EVOLUTION OF MORALITY/ III SIDGWICK'S THREE PRINCIPLES AND HARE'S UNIVERSALIZABILITY. 京都大學文學部研究紀要 1999, 38: 87-146

ISSUE DATE:

1999-03-31

URL:

http://hdl.handle.net/2433/73086

Three Essays on Ethics

Soshichi Uchii

I
UTILITY AND PREFERENCES

II
DARWIN ON THE EVOLUTION OF MORALITY

III
SIDGWICK'S THREE PRINCIPLES AND HARE'S UNIVERSALIZABILITY

I UTILITY AND PREFERENCES

1. Overview

When it comes to criticisms of Mill's utilitarianism, the distinction between the quality and the quantity of a pleasure is one of the most popular topics. Mill's statement of the distinction appears in the fifth paragraph of Chapter 2 of *Utilitarianism.*

> If I am asked what I mean by difference of quality in pleasures, or what makes one pleasure more valuable than another, merely as a pleasure, except its being greater in amount, there is but one possible answer. Of two pleasures, if there be one to which all or almost all who have experience of both give a decided preference, irrespective of any feeling of moral obligation to prefer it, that is the more desirable pleasure. If one of the two is, by those who are competently acquainted with both, placed so far above the other that they prefer it, even though knowing it to be attended with a greater amount of discontent, and would not resign it for any quantity of the other pleasure which their nature is capable of, we are justified in ascribing to the preferred enjoyment a superiority in quality so far outweighing quantity as to render it, in comparison, of small account. (ch.2, para.5)

Following this statement, Mill argues as follows: This distinction has a close relationship with the manner of human existence. Men are capable of various pleasures; but those who know well two kinds of pleasures definitely prefer the one which is obtained by employing their higher faculties. Although a being with higher faculties needs more to make him happy, and he may even experience more acute suffering

because of these faculties; but despite these liabilities, he can never wish to become a being with lower faculties. What explains this fact is a sense of dignity. All men possess this in some form, and they cannot be happy without satisfying this sense. Mill thus considered the interdependencies between pleasures, faculties, and the manner of happiness for human beings.

Now, as was pointed out by many people already, there are indeed several problems in this argument of Mill's; and I am not going to defend his distinction between the quality and the quantity of pleasures. However, despite some mistakes and confusions in Mill's argument, his argument also contains several important insights which cannot be ignored if we wish to develop a coherent theory of values on the utilitarian ground. In this paper, I wish to point out what these insights are, and to evaluate positively Mill's contributions to utilitarianism, taking also in view later development of utilitarian theories, such as Sidgwick's or contemporary writers's.

(1) First, we have to notice that Mill introduced the notion of preference into his ethical hedonism.

(2) Second, we have to notice that Mill is calling our attention, not only to the question of the quality of pleasures, but to a more fundamental question of how we make quantitative comparison of pleasures. He is treating, not only the comparison of different kinds of pleasures but also the very question of why pleasure is good and pain is evil; the latter is indeed the most fundamental question for the ethical hedonism. Let me quote Mill's words on this point.

On a question which is the best worth having of two pleasures, or which of two modes of existence is the most grateful to the feelings, apart from its moral attributes and from its consequences, the judgment of those who are qualified by knowledge of both, or if they differ, that of the majority among them, must be admitted as final. And there needs be the less hesitation to accept this judgment respecting the quality of pleasures, since there is no other tribunal to be referred to even on the question of quantity. What means are there of determining which is the acutest of two pains, or the intensest of two pleasurable sensations, except the general suffrage of those who are familiar with both ? Neither pains nor pleasures are homogeneous, and pain is always heterogeneous with pleasure. What is there to decide whether a particular pleasure is worth purchasing at the cost of a particular pain, except the feelings and judgment of the experienced ? When, therefore, those feelings and judment declare the pleasures derived from the higher faculties to be preferable in kind, apart from the question of intensity, to those of which the animal nature, disjoined from the higher faculties, is susceptible, they are entitled on the subject to the same regard. (ch.2, para.8)

(3) However, Mill tends to confound theoretical problems of hedonism with practical problems which may arise when we try to apply hedonism to our actual situations. That's the reason why his argument sometimes becomes hard to follow.

(4) Confining our attention to the theoretical problems, why is it that we can say that *the fact* that people actually prefer this to that, establishes *the value-judgment* that this is more desirable than that ? Mill's argument is not clear enough to answer this fundamental question; but it is still important in that it draws our attention to this question.

2. Preferences and a Theory of Value

Let us discuss (1) and (2) of section 1 in more detail. What is the significance of Mill's introduction of the notion of preference into the theory of ethical value? It is well known that Bentham treated the question of measurement of pleasure and pain, and he tried to establish a unified quantitative criterion for ethical values (Bentham 1789, ch.4). He tried to show how to evaluate pleasures and pains in terms of such factors as intensity, duration, certainty, propinquity, or the number of people affected. However, Bentham's discussion leaves some fundamental problems untouched.

The question of the value of a pleasure should be concerned not with the measurement of the strength of a sensation or feeling as a psychological state, but with the goodness or badness of its state; the question is evaluative, not factual. For instance, between the fatual statement "this pleasure had such and such an intensity and it lasted three minutes" and the evaluative judgment "this pleasure is good to such and such a degree", there is still a gap. You cannot infer the latter from the former by logical inference alone. Mill is certainly referring to this gap in our second quotation in the preceding section. The question of "the quantity" or "the quality" of pleasures belongs to evaluative questions, not to factual or descriptive questions; and what connect a description to an evaluation are nothing but each individual's preferences. One's preferences determine an evaluative ordering of pleasures and pains, and this ordering determines their values: to what extent they are worth having. Thus, what is essential in the utilitarian calculation is not such a bunch of factors as was mentioned by Bentham, but people's preferences: what they prefer, what they dislike. This is

what counts when we have to consider the utility; and thus we have to know what people's preferences are. As I understand, this is Mill's message when he discussed the quantity and the quality of pleasures.

I wish to add that the theory of value which Henry Sidgwick —— by far the most careful utilitarian in the 19th century —— reached was a hedonistic theory along this line, which incorporated the notion of preference into the definition of pleasures.

3. Theoretical vs. Practical Problems

Let us proceed to the third point (3) of section 1, namely the distinction between theoretical and practical problems about preferences. In ethics, as well as in science, the confusion of these two kinds of problems produces futile discussions. Mill's arguments about preferences are not entirely free from this sort of confusion, and that may well be the reason why his arguments are sometimes hard to follow. As I have already pointed out, Mill's idea that the notion of preference is indispensable for any ethical theory, whether or not it is hedonistic, is an important theoretical insight which emended the defect of Bentham's original theory; and I think Mill was basically on the right track in this. However, Mill's argument in favor of the distinction of the qualities of pleasures in terms of experts's preferences seems to me to fall largely into the practical problems of how we should apply hedonism to concrete situations. If he wants to defend the distinction between the quantity and the quality of pleasures as a theoretical distinction of hedonism, his theory may produce a contradiction. So I will argue.

Now, at the common sense level, anyone will hardly doubt that we can recognize a "qualitative" difference between one kind of pleasure and

another kind. Thus, at this level, we can agree with Mill that "it is better to be a human being dissatisfied than a pig satisfied" —— a famous dictum of Mill's. And it seems that this judgment may be paraphrased as "a man's pleasure is qualitatively higher than a pig's pleasure". However, as I have already pointed out, this judgment is not a descriptive judgment but a value-judgment, an expression of our preference. Mill is well aware of this, and he presents the condition under which this sort of judgment can claim its validity, i.e. the agreement of all who know both; in Mill's own words, "of two pleasures, if there be one to which all or almost all who have experience of both give a decided preference, . . ., that is the more desirable pleasure". However, here, it is of vital importance to distinguish the theoretical criterion from its actual applications. As I see the matter, Mill invited a number of misunderstandings and confusions by neglecting to draw this distinction clearly.

The distinction itself is simple. When Mill said that, of two pleasures, "if one of the two is, by those who are competently acquainted with both, placed so far above the other" that is so far above in its value, he was putting forward a theoretical criterion for the comparison of pleasures; and he in effect adopted the same criterion "the judgment of those who are qualified by knowledge of both" for the quantitative comparison of pleasures (see the second quoation in section 1). However, as regards two aribitrarily chosen pleasures, whether there are any who are competently acquainted with both (qualified by knowledge of both), and even if there are such people, whether their preferences are in agreement, we cannot determine a priori; these are factual problems with respect to practical applications of the preceding criterion. Further, as Mill himself was aware, if the preferences of those qualified with

knowledge of both do not agree, what should we do ? Whether or not we adopt a simple majority rule or any other, presumably supplemented by a reasonable conjecture, will also belong to the latter problem. And whether or not we can obtain a definite ordering among various kinds of pleasures is also one of the problems of practical application of the criterion, the answer of which can be obtained only after we ascertain many facts. Mill was trying to answer all these different questions almost in one breath, and that caused many difficulties.

More specifically, when he argued that "it is better to be a human being dissatisfied than a pig satisfied; better to be Socrates dissatisfied than a fool satisfied", this is really a conditional judgement which can only be justified after we ascertain a number of facts with respect to the application of the criterion (Mill just took it for granted that we know both sides and have unanimous preferences). And what is crucial for my argument is that Mill's criterion does not contain in itself anything for distinguishing the quantity and the quality of pleasures or pains. After all, he stated merely that the qualitative superiority of pleasures depends on the preferences of those who are competently acquainted with both, and he repeated the same criterion for quantitiative comparison too. Thus he didn't state anything about how we should distinguish qualitative comparison from quantitative comparison. Then, in order to make sense of Mill's criterion and argument, the only consistent interpretation seems to me to be this: the criterion for qualitative comparison is the same as that for quantitative comparison, but for the case in which, either because of a great quantitative difference, or because of an obvious and definite agreement among people, one kind of pleasure is obviously superior to another, we may conveniently and practically distinguish the two kinds as "qualitatively different". This is

in fact my own interpretation. In short, qualitative superiority of one kind of pleasure over another is a distinction at the level of practical problems, not a theoretical distinction.

4. Preferences and the Justification of Value-Judgments

Finally, let us turn to the fourth and last point ((4) of section 1). Confining our attention to the theoretical problems, how can we justify the value-judgment that pleasure A is more desirable than B, by referring to the fact that those who know both A and B prefer A to B? Unless we understand this point well, it is quite doubtful why we should appreciate Mill's insight. Let me explain how I see the matter in the following.

First of all, we have to spell out in more detail Mill's criterion of unanimous preference of those who know both sides. If one does not know well the objects of comparison or preferences, we cannot say that his preference is rational. Further, although we can hardly expect that people's preferences are unanimous on many things, it would not be so rare that different people agree in their preferences where they know quite well about the objects of preferences or expected results; on such an ideal condition, their preferences may well agree (because uncertainties, prejudices, and personal bias are excluded). In fact, on "the ideal observer" theory which has appeared occasionally in the history of ethics, this ideal observer's preferences and value-judgments are supposed to be objective.

Although the assumption of the ideal observer may be too strong as it stands, we may be able to adapt this theory to Mill's context, and we can adopt a far weaker condition. Suppose "competent acquaintance"

means that one is under no illusion, has enough information about the objects of preferences, and has actually experienced them and one's preference is rather stable; preferences under this condition may sometimes attain "objectivity", i.e. an agreement of all those with competent acqaintance. For the sake of simplicity, let us call a preference *rational* if it satisfies this condition (this notion of rationality is close to that of Brandt 1979, 10).

And this gives a clue for answering to our question of justification. People's preferences or value-judgments may well be justified if they attain this sort of agreement among those with competent acquintance; i.e. if their preferences are rational. If their preferences agree to this extent, they will also agree in their value-judgments, which means they in fact accept these judgments. This provides a theoretical condition for justifying a value-judgment. The point is that preferences satisfying certain conditions (i.e. rational preferences), not mere preferences, are the basis of justification. However, as a practical question, we are in many cases uncertain whether such conditions of preference are satisfied, and therefore we are not sure whether our value-judgment can be justified. I have no doubt that Mill was convinced that such justification of a value-judgment was possible; what he said in his discussion of the quality of pleasures confirms this. However, as I have already pointed out, Mill was not clear about the distinction between the theoretical criterion and its practical applications, and that's why his arguments are sometimes confusing and even unintelligible.

Since the problem of the justification of a value-judgment is very frequently misunderstood, I wish to add the following comments. A very popular objection based on a misunderstanding is this: "Granted that

people's preferences somehow satisfy the condition of rationality, this is a mere fact about their preferences; and how could we infer a value-judgment from this fact?" Against this objection, we have to first point out that one's value-judgment of goodness or desirability is an expression of one's preferences, not a mere description of a fact. Mill, as well as myself, would admit that there is indeed a gap between an expression of preferences and a description of facts. However, if we could ascertain the fact that people rationally prefer a certain thing and therefore they agree in their value-judgment (knowing the relevant facts), this fact is a fact about their preferences, which does not hold unless they have those preferences (thus preferences are prior). And the fact that their preferences are rational means that these preferences agree and their value-judgment also agree; thus no more can be expected for the justification. We do not infer a value-judgment from a fact; rather the fact that those preferences hold establishes the value-judgment in the sense that people in fact accept the judgment. The point of my interpretation of Mill is that his condition of "competent acquaintance" implies not merely an agreement of preferences (which may be a collective prejudice) but the rationality of preferences.

Next, one may raise the following question: "You may be able to define the condition of rational preference as you like; but the real question is whether people can satisfy the defined condition." This is a quite reasonable question. However, recall Mill argued that in several cases at least, such a condition is in fact satisfied, or at least approximately satisfied. And aside from Mill's examples, we can ourselves point out many examples of rational preferences (in our sense), if we consider calmly. For example, in any society where private property is endorsed, no one wants to be robbed of a thing dear to him.

And people with sound common sense know quite well what is "being robbed of a thing" and what it is like to them or to others; and they unanimously dislike such a situation. Thus their preferences satisfy the condition of rationality as defined above. In addition, we can enumerate many examples where the condition of rationality is at least approximately satisfied: why murder is generally wrong (imagine the underlying preferences), why a nuclear war is wrong, etc., etc. Thus, if the preceding question is meant to assert that our definition of rational preference is unrealistic, this may not apply.

Finally, someone may raise this objection (which is in fact raised by an eminent commentator against me, when I presented the original Japanese version of this paper): To try to solve the problem of justification (or any other problems for utilitarianism) by assuming an "ideal observer" just begs the question; for, even though the problem *may* be solved *if* we assume an "ideal observer", no "ideal observer" may exist, and then everything collapses and you are simply begging the question in terms of "ideal observer"!

However, this objection is another example of the confusion of a theoretical criterion and its practical application. First of all, we *did not* assume the full power of an "ideal observer" (maybe I should not have used this expression, since it invited this grave misunderstanding); we merely assumed rationality of preferences, in view of Mill's assertion. And even if we assume a stronger version of "ideal observer", this does not beg the question, because whether or not there exist someone who satisfy the condition of rational preference, or the condition of an "ideal observer", remains as an empirical question, and this is one of the essential parts of the *application* of a theoretical criterion. An ethical problem is solved if a theoretical criterion for the solution is set, *and* we

obtain a solution which *in fact satisfies this criterion.* Setting a criterion in terms of an "ideal observer" does not beg the question, since the condition is still open.

As I see it, the approach to the justification of value-judgments in terms of rational preference or rational choice is one of the major trends in the contemporary ethics (e.g., Brandt 1979, Hare 1981, Harsanyi 1977), via Henry Sidgwick's older attempt. Mill's utilitarianism, although it may have been insufficient in many respects, was a starting point of this trend, and it contained all major issues of this trend. Whether this trend is promising is still under discussion; in particular, the notion of rationality with "full information" (i.e., with some element of an "ideal observer") is sometimes questioned, and people like H. A. Simon proposed the notion of "bounded rationality" (Simon 1997) as an alternative. I should like to add that this important issue has a close bearing on the problem of justification, but that is a subject for another paper.

Bibliography

Brandt, R.B. (1979) *A Theory of the Good and the Right*, Oxford University Press, 1979.

Bentham, J. (1789) *The Principles of Morals and Legislation*, 1789.

Hare, R.M. (1981) *Moral Thinking*, Clarendon Press, 1981.

Harsanyi, J.C. (1977) *Rational Behavior and Bargaining Equilibrium in Games and Social Sciences*, Cambridge University Press, 1977.

Mill, J.S. (1863) *Utilitarianism*, 1863.

Sidgwick, H. (1874) *The Methods of Ethics*, Macmillan, 1st ed.1874; 7th ed., 1907.

Simon, H.A. (1997) *Administrative Behavior*, 4th ed., Free Press, 1997 (1st ed., 1947)

Uchii, S. (1998) "Sidgwick's Three Principles and Hare's Universalizability",

http://www.bun.kyoto-u.ac.jp/~suchii/sidg&hare__index.html

[Literature in Japanese]

Okuno, M. (1998a) "The Rationality of Action and Desire", *Annals of Ethical Studies* 28, 1998.

Okuno, M. (1998b) *Sidgwick and the Contemporary Utilitarianism*, Ph.D. thesis, Gradate School of Letters, Kyoto University, 1998.

Shionoya, Y. (1984) *The Structure of the Idea of Value*, Toyo-keizai-shinpo-sha,1984.

Uchii, S. (1988) *The Law of Freedom, the Logic of Interests*, Minerva, 1988.

Uchii, S. (1996) *Ethics and Evolutionary Theories*, Sekaishiso-sha, 1996.

II DARWIN ON THE EVOLUTION OF MORALITY

1. The Continuity of Man and Animals

Today, I wish to talk about Darwin's biological considerations on morality. There are other people who treated the same or the related problems in the 19th century, e.g. Spencer or Huxley; but it seems to me Darwin is by far the most important. When I began to study the Darwininan evolutionary theory some twenty years ago, I was very much impressed by Darwin's persistence with his thesis of *the continuity of man and animals*. In The *Descent of Man*, published in 1871 (2nd ed., 1874), this thesis is put forward as follows [Q1]:

> It has, I think, now been shewn that man and the higher animals, especially the Primates, have some few instincts in common. All have the same senses, intuitions, and sensations, — similar passions, affections, and emotions, even the more complex ones, such as jealousy, suspicion, emulation, gratitude, and magnanimity; they practise deceit and are revengeful; they are sometimes susceptible to ridicule, and even have a sense of humour; they feel wonder and curiosity; they possess the same faculties of imitation, attention, deliberation, choice, memory, imagination, the association of ideas, and reason, though in very different degrees. The individuals of the same species graduate in intellect from absolute imbecility to high excellence. They are also liable to insanity, though far less often than in the case of man. (*Descent of Man*, ch. 3)

However, traditionally, there have been various sorts of arguments for regarding man as qualitatively distinct from any other animals; among these arguments, it seems that the most persuasive was that only

man has the *moral sense* or *conscience*. For instance, Rev. Leonard Jenyns, commenting on *The Origin of Species* in a letter to Dawin, argues as follows [Q2]:

> One great difficulty to my mind in the way of your theory is the fact of the existence of Man. I was beginning to think you had entirely passed over this question, till almost in the last page I find you saying that "light will be thrown on the origin of man and his history." By this I suppose is meant that he is to be considered a modified and no doubt greatly improved orang! . . . . Neither can I easily bring myself to the idea that man's reasoning faculties and above all his moral sense could ever have been obtained from irrational progenitors, by mere natural selection — acting however gradually and for whatever length of time that may be required. This seems to me doing away altogether with the Divine Image that forms the insurmountable distinction between man and brutes. (Letter to Darwin, Jan.4, 1860. Wilson, 1970, 351.)

Thus Darwin had to face with the problem of how we can handle the moral sense within evolutionary processes, in other words, how we can give a biological explanation for man's moral faculties. This subject is tackled in chapters 4 and 5 of his book.

2. Social Instincts

Darwin's explanation of the origin of the moral sense is very interesting, but as is customary with his exposition, it is very complicated and hard to follow. But I think the main line of his argument may be reconstructed as follows: First, he puts forward the following conjecture or hypothesis [Q3]:

(H) "any animal whatever, endowed with well-marked social instincts, the parental and filial affections being here included, would inevitably acquire a moral sense or conscience, as soon as its intellectual powers had become as well, or nearly as well developed, as in man." (op. cit., ch. 4)

"Oh, come on, this is a sheer counterfactual statement, and how should we justify such a statement ?" —— no doubt many people may feel this way. But let's see what he means. Darwin means that this statement can be justified or made probable by what we know about man and social animals in general, if we supply evolutionary considerations.

First, he reminds us of a fact that man is a social animal: human beings live in a family, in a group, and in a society; and this is a biological fact like that bees and ants live in a colony. And any social animal has social instincts which support their social life. By "social instincts" he means innate or genetic propensities "to take pleasure in the society of its fellows, to feel a certain amount of sympathy with them, and to perform various services for them" (ibid.). Since social instincts are part of the "essence" of a social animal, so to speak, these instincts persist and work continually in the whole life of any individual. But these instincts may work quite differently depending on what species that animal belongs to: in the case of bees and ants, social instincts may determine particular jobs and roles an individual is to perform; but in a higher animal, social instincts may work as a mere tendency to prefer social life and to aid fellow members.

Of course, it may be asked why these animals have such instincts. Darwin has a ready answer to this: such instincts are useful for these animals, and therefore they have acquired these by natural selection. But we have to notice here that the moral sense is not included in social

instincts at this stage of the argument. Darwin's purpose is to depict the process by which the complex faculty of moral sense may be developed from the combinations of simpler faculties of social instincts and intelligence, hopefully by means of natural selection. Moreover, even if we admit his assumption that the social instincts are useful for the animals, there is still a crucial problem: useful exactly to whom ? —— to a *group* of animals or to *individual* animals ? We will come back to this problem later (Section 5).

3. Conflicts of Social Instincts with Other Instincts

Now, granted that man is a social animal, how has man aquired the moral sense ? The second stage of Darwin's argument is concerned with an imaginary psychological process which may give rise to something like moral sense or moral feeling. Suppose some social animal has acquired high intelligence so that it can remember past actions and motives. This will intensify the ability of sympathy which is included in the social instincts. Sympathy is an ability to re-present others' feelings, as well as one's own, within oneself; so that if this animal acquires better knowledge about others, by means of its improved intelligence, it is natural to suppose that the extent of sympathy will also be somehow widened.

But Darwin is *not* arguing that, since intelligence strengthens the operation of sympathy, the social instincts together with intelligence give rise to the moral sense. The matter is not that simple. We have to notice that the social instincts are not necessarily the strongest in each occasion when this animal makes decisions or actions, and they may give in to some other temporarily stronger motives, such as appetites or sexual

drive. As we all know, we humans have anti-social or selfish motives as well as social motives; we often follow the former, and with higher intelligence we may even become cleverer for satisfying our selfish motives rather than social motives. Darwin is well aware of this. Then what would give rise to the moral sense ?

The key is the enduring nature of the social instincts. The social instincts may give in to other stronger motives; but nevertheless, the social instincts are ever persistent. Then what would happen when these social instincts conflicted with other desires and were frustrated by satisfying the latter ? As we know, when a certain instinct or desire failed to be satisfied, some sort of disagreeable feeling remains. And since the social instincts are enduring, each time this animal recall this conflict, this disagreeable feeling also recurs and it may be even intensified. Thus in memory, those feelings which are associated with social instincts would become dominant. Similar things would happen with agreeable feelings of satisfaction and enjoyment; if this animal followed the social instincts rather than other desires, its satisfaction would be recalled with enjoyment, because that is quite in conformity with its enduring social nature. And this is the beginning of the formation of moral feelings; and the ability to experience these feelings is an essential part of what we call the "moral sense".

[*Note added in October 1998*: This argument was already criticized in the 19th century as trying to replace an evolutionary explanation of the origin of morality by a mere "imaginary psychology" (Shurman 1887, ch.5); and this crriticism seems to have some point. However, we can reconstruct Darwin's argument in two stages, (1) the evolution of a behavioral strategy, and (2) the evolution of psychological properties accompanying such a behavioral strategy.

As regards (1), the contemporary reader is already familiar with the

conditions under which an "altruistic" (or "conditionally altruistic") strategy can evolve and become dominant within a group. For instance, for reciprocal altruism, two conditions are necessary: (i) the same individuals must interact frequently, and (ii) they must have memory in order to respond to an opponent's previous response. We should notice that Darwin's conditions can cover these two; i.e., *social instincts* imply frequent interactions, and *intelligence* provides the memory needed for a wise strategy. I have shown, by a simple example, how a social and intelligent animal may acquire an altruistic strategy by natural selection (Uchii 1998).

As regards (2), it is quite natural to suppose that such a behavioral strategy needs some psychological makeup which supports it; in an animal with social instinct and intelligence, feelings, preferences, or propensities will accompany a behavior or a response to an opponent's action. And it is not difficult to imagine what sort of feelings are necessary for a reciprocal altruism, and this can be confirmed, to a considerable extent, by observing primates's behavior (see de Waal 1996). Thus, we can perfectly make sense of Darwin's original argument.]

4. Social Norms, Sympathy, and Habits

Darwin's emphasis on the persistent nature of the social instincts is illuminating. But his story is not over. Darwin next points out that high intelligence would be accompanied by the ability to use some sort of language, which would enable our animal to express their wishes or desires as a member of their community. Thus it is very likely that they come to form their social norms, or "public opinions" as to how they should do for the common benefit of the community. These norms or opinions are of course in an important sense "artificial" or "conventional"; and therefore these cannot be regarded as genetically determined. Darwin admits all this. But he emphasizes that "however

great weight we may attribute to public opinion, our regard for the approbation and disapprobation of our fellows depends on sympathy, which . . . forms an essential part of the social instinct, and is indeed its foundation stone" (op. cit., ch. 4). His point seems clear: although the *contents* of norms and public opinions are determined largely by artifical factors, their *binding force* essentialy depends on a biological factor, i.e. sympathetic ability, and this is instinctive or genetically determined.

The importance of sympathy has been emphasized by many philosophers such as Adam Smith or Hume. But Darwin criticizes these philosophical views as follows: we have to understand sympathy not merely as a psychological ability to reproduce former states of pain or pleasure, but also as a *biological* instinct, which is a product of evolution. Only the latter characterization can explain the fact that "sympathy is excited, in an immeasurably stronger degree, by a beloved, than by an indifferent person" (ibid.). This point is of course frequently mentioned by recent sociobiologists; but I wish to emphasize that Darwin was well aware of this, and he clearly saw its significance for ethics, although he was not clear about the biological mechanism which produces such tendencies.

By now, the major part of Darwin's view on the genesis of the moral sense or conscience has been outlined. Let me summarize his view with his own words [Q4]:

> At the moment of action, man will no doubt be apt to follow the strongest impulse; and though this may occasionally prompt him to the noblest deeds, it will more commonly lead him to gratify his own dsires at the expense of other men. But after their gratification when past and weaker impressions are judged by the ever-enduring social instinct, and by his deep regard for the

> good opinion of his fellows, retribution will surely come. He will then feel remorse, repentance, regret, or shame; . . . . . He will consequently resolve more or less firmly to act differently for the future; and this is conscience; for conscience looks backwards, and serves as a guide for the future. (op. cit., ch. 4)

In short, his explanation of the genesis of conscience has the following features: (1) it analyzes conscience into a bundle of psychological dispositions and feelings; (2) these dispositions and feelings are products of evolution and therefore are instinctive, i.e. they have a genetic basis; and (3) because of this, the workings of conscience have some conspicuous limitations that the conscience regulates mainly actions toward closer people.

The rest of his arguments is an elaboration of the preceding view. Darwin was a good observer, and it seems that this ability is well displayed in his remarks on the interplay between sympathy, public norms, and individual habits in morals. He argues that the preceding view is quite in accord with what we know about undeveloped people. Among them, only strictly social virtues are esteemed, and self-regarding virtues such as temperance or prudence are rather neglected. Darwin seems to attribute the development of self-regarding virtues mainly to the improvement of intelligence and knowledge; but he is also aware of the importance of habits of individuals. As many moral philosophers have emphasized, virtues must be acquired as a habit; and a substantial part of habits may originate from individuals and spread within their groups, and sometimes beyond their groups, by imitation. This is one of the essential features of what we call "culture". And such habits often strengthen and complement the workings of social instincts. Here,

biological process merges into cultural process. This is a very intriguing question, but we shall not get into this.

5. Darwin on Group Selection and Kin Selection

Now, what has Darwin accomplished by his argument so far ? For the sake of argument, let us suppose that his explanation of the genesis of conscience is on the right track. But where does the principle of natural selection play its role ? This still is not quite clear. Since Darwin attributed the genesis of conscience mainly to two factors, (1) intelligence and (2) the social instinct, we will examine the two in this order.

First, it seems quite clear that intelligence is developed by means of natural selection; because intelligence is no doubt useful to its possessor, an individual animal. So we can agree with Darwin's assertion, at least with respect to this factor.

But what about the social instinct ? The social instinct included sympathy, in particular, and sympathy played a crucial role in generating the moral sense or conscience. By means of sympathy, individual animals care for others and restrict their own selfish desires; in other words, *altruistic* or *moral* tendencies originate from sympathy. Then naturally we have to ask: Is the social instinct including sympathy also developed by natural selection ? Darwin's attitude to this question is ambivalent; sometimes he seems to think that the answer is obviously 'yes', but at other times he seems to be aware of a grave difficulty. But what exactly is this difficulty ? Let me explain.

Let us recall how natural selection works. There are many individual variations which are hereditary among animals of the same species. And if some of these variations are more advantageous than

others in the struggle for existence, individuals with these variations gradually increase within the species, and they eventually become dominant in number. Thus natural selection works in terms of the *herediatry* characteristics of *individuals*; and these characteristics must be useful primarily to their possessors, i.e. to individuals. But Darwin frequently speaks of moral faculties useful to a tribe or group of individuals, and he says that these faculties have been developed by *the competition among such tribes or groups* in their struggle for existence. For instance, he argues this way [Q5]: "When two tribes of primeval man, living in the same country, came into competition, if . . . the one tribe included a great number of courageous, sympathetic and faithful members, who were always ready to warn each other of danger, to aid and defend each other, this tribe would succeed better and conquer the other" (ch. 5). Granted; but is this natural selection working on individuals? Darwin doesn't seem to think it is; for he is well aware of the difficulty as follows [Q6]:

> But it may be asked, how within the limits of the same tribe did a large number of members first become endowed with these social and moral qualities, and how was the standard of excellence raised? It is extremely doubtful whether the offspring of the more sympathetic and benevolent parents, or of those who were the most faithful to their comrades, would be reared in greater numbers than the children of selfish and treacherous parents belonging to the same tribe.. . . Therefore it hardly seems probable, that the number of men gifted with such virtues, or that the standard of their excellence, could be increased through natural selection, that is, by the survival of the fittest; for we are not here speaking of one tribe being victorious over another. (ibid.)

Thus Darwin's program for explaining the genesis and development of morality by means of natural selection seems to have failed at a crucial point. That is to say, he tried to appeal to what we now call 'group selection' (i.e. an advantageous group survives and individuals of that group indirectly change), but he admitted that this group selection is not likely to be supported by natural selection working on individuals. However, it must be pointed out, to be fair to Darwin, that he was aware of at least one key for solving this difficulty. It is what we now call 'kin selection.' Just before discussing the development of moral faculties, Darwin argues for the development of intelligence by natural selection, and he briefly touches on this key, as follows [Q7]:

> If such men [i.e. intelligent men] left children to inherit their mental superiority, the chance of the birth of still more ingeneous members would be somewhat better, . . . Even if they left no children, the tribe would still include their blood relations; and it has been ascetained by agriculturarists that by preserving and breeding from the family of an animal, which when slaughtered was found to be valuable, the desired character has been obtained. (ibid.)

This idea could have been more developed and applied to the explanation of moral faculties; but Darwin left that job to the biologists in the 20th century, such as W. D. Hamilton (kin selection) or Robert Trivers (reciprocal altruism). What Darwin actually did instead was to appeal to the principle of heredity of acquired characters.

6. The Significance of Darwin's Considerations on Morality

In this talk I have outlined what I take as the essence of Darwin's

theory of morality. He was mainly concerned with the biological and psychological task of explaining the genesis of moral faculties of man. But it seems to me that he was also interested in moral philosophy based on the evolutionary theory. The major advocate of what is called 'evolutionary ethics' in the 19th century was of course Herbert Spencer; and Darwin was far more cautious than Spencer, trying to avoid any definite statements about what we *ought* to do. But now and then he criticizes former and contemporary moral philosophers, such as Adam Smith or John Stuart Mill, and sometimes even gets into issues of eugenics, in *The Descent of Man.* This indicates Darwin's strong interest in moral philosophy. Moreover, we have good evidence that this interest originates in his youth. For instance, I was surprised by finding the following remarks (written in October, 1838) in his Notebooks [Q8]:

> Two classes of moralists: one says our rule of life is what will produce the greatest happiness. —— The other says we have a moral sense —— But my view unites both & shows them to be almost identical. What has produced the greatest good or rather what was necessary for good at all is the instinctive moral senses: (& this alone explains why our moral sense points to revenge). In judging of the rule of happiness we must look far forward & to the general action —— certainly because it is the result of what has generally been best for our good far back. —— (much further than we can look forward: hence our rule may sometimes be hard to tell). Society could not go on except for the moral sense, any more than a hive of Bees without their instincts. (Old & Useless Notes 30, Barret et al., 1987, 609.)

We may recall that the moral philosophers who emphasize the moral sense are called 'Intuitionists' and those who emphasize the greatest happiness are called 'Utilitarians'. Thus the young Darwin here is

claiming that he can synthesize these two major schools of moral philosophy! I will add, for your curiosity, that Henry Sidgwick, a great utilitarian and who claimed that Intuitionism and Utilitarianism can coincide, was born in the same year, 1838. And we have to notice also that Darwin's idea of the genesis of morality is already sketched in rough outline in the last sentence.

But these historical interests aside, are there any significant suggestions for ethics or normative moral philosophy that can be exploited from Darwin's theory of the moral sense? I think there are. Since I do not have much time left, let me briefly touch on this without arguments. First of all, (1) we have to know well about human morality in order to make any normative assertions. And in this respect, the Darwininan view of morality is certainly useful. We have to construct feasible ethics for humans as a social animal, not for an angel or an isolated beast. For this purpose, we certainly have to know our biological capacity, limitations as well as potentialities.

Secondly, (2) if the Darwinian view is on the right track, we should take the continuity of man and animals more seriously. Darwin argued more or less persuasively that we humans and other animals share many properties, including intelligence, feelings and preferences. Hence, if we find some of these valuable and think that they should be protected by our morals, the same consideration should support similar treatments of animals, with the difference of various degrees, of course. For instance, persons like Jane Goodall, knowing very well about primates, assert that our treatment for them should be improved; and this assertion may well be justified.

Thirdly and finally, (3) the Darwinian view suggests a certain approach to ethics, say the Reductionist approach (I borrow this word

from Parfit, who uses it in the context of the problem of personal identity; and Daniel Dennett also defends this approach, with respect to cognitive science, in his *Darwin's Dangerous Idea*, 1995). This is the view that all ethical concepts can be analyzed into more basic concepts which are not themselves ethical. In other words, it is the view that concepts such as 'conscience' or 'moral goodness' will be well understood only in terms of concrete workings of human faculties and feelings, without postulating any peculiar realm of moral value. This is exactly what Darwin has done in his theory of the moral sense; conscience or moral sense is so called because of its workings in a certain way, not because it is related to some irreducible moral value. Since this position is very likely to be misunderstood, I will hasten to add a few explanatory remarks.

By reductionism I do not mean that ethical or evaluative concepts can be reduced to factual or descriptive concepts; this is what Moore called 'naturalism' and I do not support it. In order to be a reductionist in my sense, one need not be a naturalist. All one has to admit as an ethical reductionist is that morality can be related to a bunch of natural or conventional elements and their workings. Morality needs intelligence, but this intelligence does not come from any peculiar realm, devine or angelic. Morality needs some instinctive factors, but one can find similar factors in other animals. And, again, moral feelings and preferences have an origin in a non-moral animal world, and you don't have to suppose any peculiar 'respect for the divine moral law'. All the factors necessary for full understanding of morality can be found in this world and the workings of its constituent parts. This is what I mean by reductionism.

And I understand that Darwin is one of the most powerful advocates of this position, although very few people would regard him as a moral

philosopher. So, by emphasizing his contribution to ethics as a reductionist, I should like to end my talk.

Bibliography

Barret, P.H. et al., eds. (1987) *Charles Darwin's Notebooks, 1836-1844*, Cambridge: Cambridge University Press, 1987.

Darwin, C. (1871) *The Descent of Man*, 1st ed., London: Murray, 1871.

Darwin, C. (1874) *The Descent of Man*, 2nd ed., London: Murray, 1874. (1922 reprint is used.)

Dennett, D.C. (1995) *Darwin's Dangerous Idea*, Simon and Shuster, 1995.

de Waal, Frans (1996) *Good Natured*, Harvard University Press, 1996.

Parfit, D. (1984) *Reasons and Persons*, Oxford University Press, 1984.

Shurman, J.G. (1887) *The Ethical Import of Darwinism*, Charles Scribner's Sons, 1887; 3rd. ed., 1903.

Uchii, S. (1996) *Evolutionary Theory and Ethics* [in Japanese], Kyoto: Sekaishiso-sha, 1996.

Uchii, S. (1997) "The Origin of Morality" [in Japanese], *Kagaku* (*Science Journal*) 67-4, 1997.

Uchii, S. (1998) "From the Origin of Morality to the Evolutionary Ethics, part I" [in Japanese], *Tetsugaku Kenkyu* (*Journal of Philosophical Studies*) 566, October 1998.

Wilson, L.G., ed. (1970) *Sir Charles Lyell's Scientific Journals on the Species Question*, New Haven: Yale University Press, 1970.

Postscript

This is a paper read for the session on 19th Century Biology, International Fellows Conference (Center for Philosophy of Science, Univ. of Pittsburgh), May 20-24, 1996, Castiglioncello, Italy. Robert Butts was the commentator; his comments and questions from the floor mostly centered on what I *didn't say* in the paper, i.e. on the point how scientific knowledge of evolution and normative ethics are related. I have worked out my ideas in my book (1996, see the preceding Bibliography); the essential idea is that (1) evolutionarty biology can teach us what sort of sentiments

or preferences we have as part of our human nature, and (2) moral sentiments and preferences are among them. Since, as I see it, the justification of moral judgements can be made essentially in terms of our rational choice for satisfying our preferences (not all, but those that can survive criticisms by facts and logic)—— including moral preferences——, evolutionary knowledge, unlike knowlege of general relativity or quantum mechanics, does contribute to our normative ethics. For a brief discussion of the justification of "ought" statements (prudential, moral, etc.), see my paper "Comments on Prof. Ruse's View" in *PHS Newsletter,* No.19, Nov. 1997 (http://www.bun.kyoto-u.ac.jp/phisci/Newsletters/newslet_19.html).

Finally, I wish to thank Jerry Massey for teaching me de Waal's recent book (1996) on the origin of morality; I have supported my view by de Waal's observations in my 1997 and 1998 papers.

Ⅲ SIDGWICK'S THREE PRINCIPLES AND HARE'S UNIVERSALIZABILITY

In this paper, I wish to draw the reader's attention to certain similarities between Sidgwick's and Hare's view on what is called the 'universalizability of a moral judgment'; and, further, I wish to show that, despite these similarities, there are some important differences between them. While Sidgwick's principles may be all regarded as a kind of impartiality, Sidgwick insisted they are non-tautological; whereas Hare's universalizability is meant to be a logical thesis established on logical grounds. Contrary to our initial prejudice that Hare is clearer than Sidgwick in many respects, it will turn out that these differences show that Sidgwick's analysis is deeper than Hare's. I will substantiate this claim by showing that Hare's theory of critical thinking makes use of the evaluative principles corresponding to Sidgwick's three principles.

1. Sidgwick's Three Principles

It is well known that Henry Sidgwick propounded a version of utilitarianism based on the three self-evident principles and the hedonistic theory of the ultimate good. The three principles are:

(1) *The Principle of Justice*: this constrains the judgment of 'right' or 'ought' as follows: "whatever action any of us judges to be right for himself, he implicitly judges to be right for all similar persons in similar circumstances" (Sidgwick 1907, 379).

(2) *The Principle of Prudence*: this is related to the notion of the good

on the whole of a single individual, and is stated as follows. "Hereafter *as such* is to be regarded neither less nor more than Now"; "the mere difference of priority and posteriority in time is not a reasonable ground for having more regard to the consciousness of one moment than to that of another" (381).

(3) *The Principle of Rational Benevolence*: this is about the universal good, i.e. the good of all individuals, and this principle is composed of two propositions: "the good of any one individual is of no more importance, from the point of view of the Universe, than the good of any other"; so that "as a rational being I am bound to aim at good generally, —— so far as it is attainable by my efforts, —— not merely at a particular part of it" (382; I prefer this formulation in this paper, because the two components are stated separately).

One may notice that all may be regarded as a kind of impartiality, the object of impartial treatment being different in each case. However, we have to be careful. In stating these three principles, Sidwick is insisting that all the three are *non-tautological*, i.e. not provable on logical grounds alone and have some substantive content; and, further, that (1) is a substantive principle obtained from the consideration of a *Logical Whole*, whereas (2) and (3) are a substantive principle obtained from the consideration of a *Mathematical or Quantitative Whole* (380-381). It is of vital importance for our interpretation of Sidgwick's three principles that we understand the exact implications of this assertion.

2. Hare's Analysis of the Universalizability

On the other hand, it is also well known that Richard M. Hare has

propounded another version of utilitarianism based on the logical properties of moral words and the requirement of rationality. The logical properties of moral words he mainly appeals to are: the universalizability and the prescriptivity of a value-judgement. Since the main point of this paper is concerned with the affinity and the difference between Sidgwick's three principles and the applications of Hare's universalizability, I shall ignore the prescriptivity, and concentrate only on the universalizability. Now, what is the universalizability of a value-judgment ?

Hare gives, as we shall shortly see, several kinds of explanation, but the essential content of the universalizability seems to be clear. He has been maintaining "that the meaning of the word 'ought' and other moral words is such that a person who uses them commits himself thereby to a universal rule" (Hare 1963, 30). 'Universal' means that it does not contain any reference to an individual, such as a particular person, a particular time, or a particular spatial location. Thus, according to Hare, an 'ought'-judgment like "He ought not to smoke in this compartment" —— although it is a singular judgment referring to an individual person 'he' (whoever it is) ——, depends on, or implies, another 'ought'-judgment which does not contain any reference to an individual, and hence can be expressed only in terms of universal quantifiers and universal words.

But why does the word 'ought' or any other moral (evaluative) word have this property of the universalizability? Hare gives the following reasons (roughly in a chronological order, as Hare's view develops).

(1) First, an 'ought'-judgment (and a value-judgement in general) must be made on a *criterion*; and this implies that if the same criterion is satisfied, the same judgment must be made. Thus as long as a value-judgment is made on a criterion, it is implicitly universal (Hare 1952,

ch.6).

(2) Second, an 'ought'-judgment (and a value-judgement in general) must be supported by a *reason*; and this implies that the same judgment must be made whenever the same reason holds. Thus there must be a universal 'ought'-judgment behind the individual 'ought'-judgment (Hare 1952, 176; Hare 1963, 21, etc.).

(3) Third, an 'ought'-judgment (and a value-judgement in general) has a *descriptive meaning*; and a descriptive meaning presupposes a universal rule which determines it. Although evaluative words and descriptive words differ in their essential function, they do share this feature as long as they have a descriptive meaning as an element of their meaning (Hare 1963, ch.2). And since the meaning-rule which governs the use of a descriptive term is a universal rule (dependent on the *similaritiy* of objects in a certain respect), a singular descriptive judgment is universalizable, and in the same way, a singular 'ought'-judgment (and an evaluative judgment in general) is universalizable (Hare 1963, 13).

These are the major lines of arguments in favor of the universalizability. But our question is: are they right, or is any of the three reasons good enough to establish the universalizability ?

3. Is the Universalizability True on Logical Grounds ?

Let us recall that Hare is trying to show the universalizability as a *logical* thesis: it is meant to be true on conceptual grounds alone, or true by virtue of the meaning of 'ought' (or any other evaluative word). Thus, if Hare is right, the universalizability is analytically true, given the meaning of 'ought'; or it can be established by the conceptual truth contained in the notion of 'criterion' (sect.2, (1)) or 'reason' (sect.2, (2)) or

'meaning-rule for a descriptive word' (sect.2, (3)).

However, we also have to recall that Sidgwick insisted that all of his three principles are *non-tautological* (see Okuno 1998, 6.2). One of the easiest interpretation of his Principle of Justice (sect. 1, (1)), for instance, is that it merely expresses the universalizability of 'ought' or 'right action' (I myself endorsed this interpretation for sometime). But if Hare is right, this interpretation makes the Principle of Justice a tautologous or analytically true principle, and Sidgwick persistently tried to avoid such a principle for his 'self-evident' principles for ethics. Then it is obvious that both cannot be right; so which is wrong, Sidgwick or Hare ?

Actually, I once argued (a long time ago, Uchii 1974) that Hare is wrong; and although I have not changed my mind, I now see the reason more clearly why he is wrong, because I have now realized the important differences between Sidgwick's three principles. Let us first concentrate on the universalizability of 'ought'.

No one will deny that an 'ought'-judgment has a criterion. No one will deny that an 'ought'-judgment must be made on some reason. And, again, no one will deny that an 'ought'-judgment has a descriptive meaning. However, it is not so obvious that the criterion must be universal, in the sense that it does not contain any reference to an individual. Likewise, it is not so obvious that the reason must be universal in the same sense. And, finally, it is not so obvious that the descriptive meaning must be explicable without any reference to individuals.

Since Hare seems to be emphasizing the universalizability as the common property between descriptive judgments and value-judgments (in Hare 1963), let us consider the descriptive meaning of a descriptive judgment, such as

(1) This is one meter long.

I presume no one will question that this is a descriptive judgment. Indeed, Hare (1955) mentioned 'one meter long' as an example of a universal expression (Hare 1955, 306). But how can we explain the descriptive meaning of 'one meter long'? You know that the unit of the length 'meter' was determined historically, and there is the standard of 'meter' in Paris. Thus, 'this is one meter long' means that the length of this is identical or at least approximately identical with that standard; more specifically, it would mean that 'if this is transported to Paris and compared with the standard, the two will coincide'. As Hare rightly points out, any descriptive word depends on the similarity or the identity of this sort, and we can say that

(2) *anything* similar to the Paris standard in the relevant respect (i.e. the length) is 'one meter long'.

This statement certainly has a universal form, and it defines an open class. But the *syntactical form* of universality is only a necessary condition for the universalizability; and likewise an open class may not correspond to the extension of a universal word, since you can define an open class by referring to an individual, as is the case with 'a citizen of the United States'.

Notice that, with the last example, we can construct a similar statement to (2) as follows:

(3) *anyone* similar to John F. Kennedy in the relevant respect (i.e. nationality) is 'a citizen of the United States'.

This statement also has a universal form, but no one will deny that it (implicitly) contains an essential reference to an individual (the United States). Thus it is clear a statement with a syntactically universal form may not be properly or *semantically* universal.

And (2) is such an instance as containing a reference to an individual. I know Hare claimed that the expression 'similar to X' can be replaced with a universal word (Hare 1955, 306-7; 1963, 11) even if X is a singular expression; but his claim is without a proof, and refuted by our example (3). May I also point out that the Paris standard is a *unique individual*? Now, can we eliminate from (2) the reference to the Paris standard? You might think that the reference is inessential because you can substitute a reference to another standard (many countries have their own copies of the standard); but you must recall that such substitute standards work as a standard precisely because of their connection with the Paris standard; and (2) is an instance of what Reichenbach called a 'coordinative definition', a definition correlating a concept to a particular object (Reichenbach 1958, 14*).

---

*Reichenbach 1958, 14:

Physical knowledge is characterized by the fact that concepts are not only defined by other concepts, but are also coordinated to real objects. This coordination cannot be replaced by an explanation of meanings, it simply states that *this concept* is coordinated to *this particular thing*. . . . these first coordinations are therefore definitions which we shall call *coordinative definitions*.

I know that more recent methods of determining a meter are more complicated and refer to the wave length of a spectrum of a certain atom or of light; still, unless these complicated methods retain their reference to the original standard, the meaning of 'meter' will change. But we do not have to get into messy details. The point here is that the meaning-rule which determines the meaning of 'meter' did historically contained a reference to an individual, and nothing was wrong, logically, with this meaning-rule; statement like (2) is not semantically universal, because it contains an essential reference to an individual, but it works perfectly well as a rule for determining the meaning of a descriptive word. This is a *convention* indispensable for *making* 'meter' a universal word; thus a meaning-rule for a descriptive word is not as simple as Hare supposed.

4. The Weak and the Strong Universalizability

Let me distinguish the *weak universalizability* from the *strong universalizability* (although the words sound very similar, my distinction is quite different from Gibbard's 1988, 59-60* ; as I see it, his distinction is rather concerned with *weights* of preferences, which will be discussed in the subsequent sections 5, 6, and 7). I will agree with Hare that a descriptive or evaluative word (having a descriptive meaning) is universalizable in the form of (2), and I will call this the weak universalizability. And if we can eliminate all words containing a reference to an individual (e.g., the Paris standard), or if we can replace all such words with properly or semantically universal words (i.e., words with no reference to individuals), I will call this the strong universalizability. Then, my point can be expressed in a word; namely, the logic of a descriptive word does not necessarily demand the strong

universalizability.

---

*Gibbard 1988, 59-60:

A moral statement, Hare says, is an overriding prescription that is universalizable: the prescriber must stand ready to prescribe the same thing no matter what position he is to occupy. . . .

Weak universality requires only that I prefer all told the same alternative for any position I might occupy. It does not require that my preferences all told be equally strong for each of those positions. I can care what position I occupy, so long as I do not care enough to reverse the direction of my preferences all told. If, on the other hand, a person's preferences all told are position-independent in strength as well as direction, then I shall call them *strongly universal.*

If we grasp this point, the rest of my argument is quite easy. The existence of a criterion for a descriptive or evaluative word does imply the weak universalizability, but not the strong universalizability, because the criterion may contain a reference to an individual. The existence of a reason for a value-judgment does imply the weak universalizability, but not the strong universalizability, because the reason may contain a reference to an individual. Thus, if we wish to assert the strong universalizability of a value-judgment, we need *more* than the logic of a descriptive meaning, *more* than the logic of a criterion, *more* than the logic of a reason (I pointed this out in Uchii 1974, but I did not know Sidgwick well then). Thus, although Sidwick may not have known the modern logic, his intuition was quite acute. When he asserted his Principle of Justice is not tautologous, he was basically right. The strong

universalizability of 'ought' or 'right' has some substantive content, not provable by logic alone.

There remain the following questions: Then, (a) is 'ought' universalizable in the strong sense, (b) and if it is why ? But I will not pursue these questions here. All I wish to point out is that if you want to give an affirmative answer to the first question (a), you have to defend your answer on a stronger assumption than Hare's; the fact that many people admit the strong universalizability of 'ought' does not establish that it is a *logical* thesis. Some may wish to appeal to the concept of morality (e.g., by asserting that at least '*moral* ought' is universalizable), and others may admit that the strong universalizability (with respect to evaluative words) is itself a substantive ethical principle, despite its formal and abstract character. But in either case, its justification is needed. Notice that, even if we make the universalizability true by vitrue of the meaning of 'moral', we thereby import another substantive question, 'why should we be moral ?' Thus, although many of us are , unlike Sidgwick, unhappy with an appeal to 'self-evidence', Sidgwick's claim that the (strong) universalizability of 'ought' is non-tantologous seems still correct.

5. Universalizability and the Concept of Good

Let us get back to Sidgwick's Principles. In addition to the Principle of Justice, Sidgwick mentioned two other, i.e. the Principles of Prudence and of Rational Benevolence; and he claimed that none of them are tautologous. Whereas Hare seems to have derived, in effect, in his *Moral Thinking* (1981) by means of the logic (the prescriptivity and universalizability of an evaluative judgement) and the facts of the case in

question what these Principles have accomplished; thus leading to his version of utilitarianism. But as we have already seen in the last section, the strong universalizability (Hare clearly subscribes to this) is not tautologous (analytic) and has some substantive content, and in this respect Sidgwick was right. This raises a strong doubt about the validity of Hare's 'derivation' of his own version of utilitarianism. But we will first examine Sidgwick's view.

To begin with, why does Sidgwick need two more Principles in order to give the basis of utilitarianism? Let me quote one of the relevant passages from *The Methods of Ethics* (Sidgwick 1907) at length, because this is very important:

> The principle just discussed, which seems to be more or less clearly implied in the common notion of 'fairness' or 'equity', is obtained by considering the similarity of the individuals that make up a Logical Whole or Genus. There are others, no less important, which emerge in the consideration of the similar parts of a Mathematical or Quantitative Whole. Such a Whole is presented in the common notion of the Good —— or, as is sometimes said, 'good on the whole' —— of any individual human being. (380-1)

Sidgwick is first trying to explain why the Principle of Justice (roughly, the strong universalizability of 'ought') is not tautologous. His reason is not quite clear, but he seems to be suggesting that, although it is logically possible to treat differently different individuals making up a Logical Whole (humans, in this case), our reason dictates to treat them similarly, if their situations are similar; and that this dictate is self-evident, although it is non-tautologous. Then, Sidgwick turns his attention to another kind of Whole (called Mathematical or Quantitative Whole), and points out that one's Good on the Whole is such a

Quantitative Whole. He then continues:

> The proposition 'that one ought to aim at one's own good' is sometimes given as the maxim of Rational Self-Love or Prudence: but as so stated it does not clearly avoid tautology; since we may define 'good' as 'what one ought to aim at.' If, however, we say 'one's good on the whole', the addition suggests a principle which, when explicitly stated, is, at any rate, not tautological. (381)

It should be clear that Sidgwick is carefully trying to avoid a tautologous principle. And recall that the Principle of Prudence is stated as: "Hereafter *as such* is to be regarded neither less nor more than Now". This prescribes how we ought to treat different parts of one's good on the whole; and Sidgwick is pointing out that the last notion is not a Logical Whole but a Quantitative Whole. Many readers may be puzzled by this distinction; is 'good' so different from 'ought' or 'right'? Yes, it is, and I will explain the difference on behalf of Sidgwick (judicious Schneewind 1977 is not of much help on this point; see 298-300).

Whether or not an act is right, whether or not you ought to do it, is a two-valued distinction; there is no middle-road option, such as 'this act is a-half right', so that it is not a quantitative distinction which allows a difference of degree. And Sidgwick is saying, in the Principle of Justice, that such a distinction should equally apply to any two individuals similar in the relevant respect. On the other hand, whether something is good for me is clearly a matter of degree, 'good' being essentially a matter of comparison; moreover, Sidgwick is committed to the view that one's good on the whole must be composed of one's particular good experienced at each moment. Sidgwick distinguishes 'ultimate good' from 'good as a means' and he is talking here about the former. By introducing

the notion of 'one's good on the whole', Sidgwick is drawing our attention to the relation of parts to the whole and the relation of parts to other parts. Notice that such a relation brings in new problems which do not arise in the rightness of an action: *how should one compare one good to another*, and *how should one reflect the comparative value of one good into the value of the whole good*?

This problem is quite independent from the (strong) universalizability of 'ought' or rightness. To see this, you need only to imagine the following sort of cases: Suppose you assign a higher value to any of your particular goods according to the closeness of them to the present moment, *now*. Although this is quite contrary to the Principle of Prudence, this choice (conceived as an action) can satisfy the universalizability, as long as you continue to prescribe consistently the same choice to yourself (in earlier and later moments) and others: 'Since this choice is right for me now, it is right for me at other moments, and for anyone at any moment' (notice that 'now' in this judgment can be expressed by a universally quantified time variable). Of course such a choice makes the determination of the value of good on the whole awfully difficult (if possible at all, and you may need another principle for summation); but again, this has nothing to do with the universalizability of rightness.

For the sake of comparison, let us examine the same problem from Hare's standpoint. Hare applies the universalizability to goodness (and to any other evaluative concept). But we have to be careful not to overestimate the extent of its application, in view of Sidgwick's analysis. For a while we will ignore the distinction between 'ultimate good, or good in itself' from 'good as a means', since this distinction is not central

in Hare. Now, if I say

(1) this is a good philosophy book,

I am committed to

(2) any philosophy book similar to this in the relevant respect is good,

according to the (strong) universalizability; although I think the (strong) universalizability applied to goodness is still non-tautologous if 'the relevant respect' is taken to be universal, I shall ignore this point. Now, if 'good' is a *comparative* notion (no one will deny this), and if it is further a *quantitative* notion (comparability does not necessarily imply this, because a mere ordering is insufficient for producing a measure of goodness), the universalizability is quite incompetent to impose any restriction on such a comparison or a quantitative measure.

Soppose I wish to make a ranking list of philosophy books I have ever read. Does (2) impose any restriction on such a ranking? Yes, it does, but very little. For, whatever criterion I may be adopting for evaluating philosophy books, (2) implies merely that if the same criterion is satisfied, I have to call a book 'good'; it does not teach *where* I should insert that book in my ranking list. If you think (2) can do more, you are implicitly adding something more to the universalizability. For instance, it may be supposed that the criterion for the evaluation does provide a clue how to *grade* a philosophy book; but the universalizability merely says 'the same condition, the same grade', and does not tell anything about *how* I shoud *grade*—— this already presupposes a sort of

comparative or *quantitative* notion of goodness. Thus, it seems this consideration confirms Sidgwick's distinction. Any criterion of goodness, since 'good' is a comparative or often quantitative notion, must refer to a method of ordering or grading, in addition to specifying the relevant respects for evaluation. 'Ought' and 'right' need only the latter.

Hare may be able to bring in *preferences* for laying down a criterion of comparison of goodness or a (quantitative) measure of goodness. However, it is not clear at all how the (strong) universalizability may help for determining such a criterion based on preferences. In particular, when we have to determine the goodness of a book I read *in the past* in comparison with another book I read *now*, the preceding problem of comparing a past good (preference) with a present good (preference) appears in Hare too. That is exactly the reason why he avoided discussing the problem of the 'pure discounting of the future' (i.e., giving less weight to future preferences; see Hare 1981, 100-101). If the universalizability can solve this problem, why didn't Hare do that? And if Hare claims pure discounting is *irrational*, Hare is not different from Sidgwick.

6. Universalizability and Benevolence

Next, let us turn our attention to the Principle of Rational Benevolence. While the Principle of Prudence is related to one's good on the whole, Rational Benevolence is related to the good of *all* individuals (taken together). Its point is that a person's good should be treated equally with another person's good, if their amount is the same. That this Principle is independent of Prudence is clear, since the equality of weight *through time* in *one* individual's good does not say anything about weights

of *diffent individuals* when we have to consider their good taken together. Sidgwick says as follows:

> And here again, just as in the former case, by considering the relation of the integrant parts to the whole and to each other, I obtain the self-evident principle that the good of any one individual is of no more importance, from the point of view (if I may say so) of the Universe, than the good of any other; unless, that is, there are special grounds for believing that more good is likely to be realised in the one case than in the other. (Sidgwick 1907, 382)

It is clear that Sidgwick is appealing to the notion of a Mathematical or Qunatitative Whole, this time that of the universal good (on the whole). It should be clear by now that his Principle of Benevolence is also independent from that of Justice or the strong universalizability, and further, is non-tautologous. For, the universalizability has nothing to do with a quantitative measure, and it is logically quite possible that someone's good is weighted twice as much as another's good; notice that if you take the Egoistic method in Sidgwick's sense, you are giving a dominant weight to your own good. But Sidgwick is saying that rationality demands, *if we take the point of view of the Universe,* to give equal weight to everyone's good; and thus the Principle is non-tautologous. This conditional character of the Principle of Benevolence is amply made clear by Okuno (1998b, 7.1.3, 7.3), and because of this conditional character, this Principle is consistent with that of Prudence (thus my comment on these principles in Uchii 1988, 220 is wrong).

Although Sidgwick's distinction between a Logical Whole and a Quantitative Whole appeared somehow abruptly and its significance was not so clear initially, it thus turned out that its importance is great for

ethics. I must confess that although I knew Sidgwick these twenty years, I have been trying to interpret him mainly in terms of Hare's universalizability. J. B. Schneewind likewise asserts that all the inferences required by the three principles are generalizing inferences, the reason thereby prohibiting us to take arbitrary differences into consideration (1977, 300-302). Another interpretation in terms of application of fairness appears in Shionoya (1984, 156-7; he confounds a Quantitative Whole with a Logical Whole). However, all such attempts miss the real significance of Sidgwick's distinction between Logical and Quantitative Whole. I realized this only last year, as a byproduct of my study of the philosophy of space and time and reading some of Harsanyi's papers on social-welfare function (Harsanyi 1976, 1977, 1982). Space-time philosophy suggested the conventionality of simultaneity and the geometrical structure; given a space-time manifold, it is still a long way to determine its metric structure and we need to introduce many assumptions such as the unit of length, the method of measurements, the definition of simulaneity, etc. Similar things may well happen in ethics; e.g., given one's goods at particular moments, we still need a principle to form one's good on the whole; and likewise, given the good of each individual, we still need a principle to form the good of all individuals taken together. Harsanyi has shown, more technically, how to do this, although there still remain difficult problems in order to reach the usual maximization principle of the sum of individual utilities. Sidgwick was addressing himself to this sort of problem.

In order to illustrate further the importance of Sidgwick's distinction, I will criticize in the next section Hare's 'derivation' of utilitarianism, in the light of Sidgwick's insights.

7. Hare's Implicit Use of Evaluative Principles

I will first summarize briefly Hare's 'derivation' of a utilitarian conclusion from the facts of any given case via the logic of evaluative words, in his *Moral Thinking* (1981). I put the word 'derivation' within quotes because it is not a straightforward logical inference from premisses to a conclusion; rather, his 'derivation' means that if one (1) tries to *decide rationally* what one *ought* to do in the given case, (2) knowing all the *facts* and (3) following out the *logic of moral judgments*, then one will *accept* a certain 'ought'-judgment, and (4) this 'ought' prescribes an act which *maximaizes the satisfaction of preferences of all persons* involved in the case. As is well known, Hare distinguishes the critical thinking from the intuitive thinking in moral discourse, and I am here talking only about the critical thinking (which assumes full rationality and sufficient information; I will touch upon the problem of rationality later). Despite the criticism I am going to put forward in the following, I still appreciate Hare's method for justifying an evaluative conclusion in this manner, leaving the gap between Is (description) and Ougut (universalizable prescription) as it is.

Let us begin with his simplest model case, the car-bicycle example (Hare 1981, 6.2). Adam wants to park his car but Eve has put her bicycle in the only vacant parking lot; Adam prefers to park, whereas Eve prefers her bicycle to stay where it is, but it is assumed that Adam's preference is stronger than Eve's (this presupposes the interpersonal comparison of preferences). Assuming, further, that Adam has perfect knowledge about all this, and he is ready to decide what he *ought* to do by critical thinking, what ought he to do? Hare answers as follows:

Since Adam wants to decide what he *ought* to do, by

universalizability he must seek an 'ought'-judgment which is acceptable even if Adam's position and Eve's position are reversed; thus if he is to accept

(a) Eve ought to move her bicycle in order to enable me to park there,

he must be ready to accept

(b) If I were in her position, I ought to move my bicycle.

But for this, *he must know what it is to be in Eve's position with her preferences*, because the proposed act will frustrate some of her preferences; and because a *rational decision* must be made in the light of knowledge and logic. Further, this knowledge in the context of critical thinking satisfies what Gibbard (1988, 58) named the *Conditional Reflection Principle*: 'I cannot know the extent and quality of others's sufferings and, in general, motivations and preferences without having equal motivations with regard to what should happen to me, were I in their places, with their motivations and preferences' (Hare 1981, 99). Hare regards this as a conceptual truth (by virtue of 'know' in the moral context). In short, Adam has to represent Eve's preference for unmoved bicycle within himself by *his own acquired preference* equal in strength with hers; this preference is a consequence of his knowledge, by the Conditional Reflection Principle.

Then, the problem for Adam is now reduced to a *rational decision as regards his own conflicting preferences*. And since if his own two preferences conflict the stronger wins, a rational or prudential choice is

to satisfy the stronger (here, 'rational' roughly means that the choice survives the criticism in the light of logic and sufficient knowledge). And the same solution applies in our case too: since Adam's original preference is assumed to be stronger than Eve's preference now represented by Adam's acquired preference, the rational solution is to choose an 'ought'-judgment which satisfies (maximaizes the satisfaction of) Adam's overall preference (everything considered), in this case (a). Thus the 'ought'-judgment is justified in terms of a rational acceptance based on facts and logic.

This solution seems very attractive. Unlike Sidgwick, Hare seems to have dispensed with the notion of 'one's good on the whole' or of 'people's good on the whole', thereby avoiding Sidgwick's 'Quantitative Whole'. However, on a closer examination, similar problems reappear (I was still unaware of this in my 1994). First, even in the case of intrapersonal comparison of one's own preferences, the comparison is not always among contemporaneous preferences, often involving conflicting preferences ranging over different times; the question of prudence will lose much significance if we may restrict our attention only to contemporaneous preferences! What is mostly at issue is the problem of *diachronic* rationality, requiring the comparison of preferences at different times. Hare is of course aware of this problem; that's why he spent many pages (Hare 1981, 101-106, 124) for discussing 'now-for-now' and 'now-for-then' and 'then-for-then' preferences, emphasizing that 'there are obvious analogies between other people's preferences and our own preferences in the future' (124). This clearly shows that, although Hare did not use the notion of 'one's good at a moment', Sidgwick's problem for prudence (rational self-love) reappears in a different form in

Hare's theory too. And in view of Sidgwick's discussion, the essential problem is: how to *weigh the importance* of preferences at each moment?

And this directly leads to the second point. We also have the corresponding problem related to benevolence: how to weigh *the importance* of one individual's preferences against another individual's preferences? These problems are analogous but independent, as Sidgwick clearly saw. But Hare seems to think these problems can be somehow solved in terms of universalizability, which I now realize is wrong. Although I was quite impressed by Hare's preceding argument, I always felt uneasy about his step from particular preferences to the final preference ('prefertence all told') for choosing an 'ought'-judgment. There is the problem of 'correct representation' of another person's preferences (or of one's own at different times), in the first place. This problem seems quite analogous to the measurement of length, which has to postulate a *unit of length* and a definition of *congruence at different locations*; it does not make sense to postulate the existence of the *absolute length*, independent from these postulates. In our preference case, we have to have, likewise, criterion for telling whether this preference is the same as, or greater than, another preference, in strength, if two are not contemporaneous within oneself, or if two belongs to different individuals. Intrapersonal comparison, as well as interpersonal comparison, depends on such a criterion. Hare is aware of this, when he says: 'We imaginatively suppose that we could have the choice of having one of these experiences or the other, and from a preference *now*' (Hare 1981, 125); thus introducing the present (informed) preference as the criterion of the comparison (for a clearest statement of Hare's problem, see Giffin 1988, 76*).

*Griffin 1988, 76:

Suppose I know a lot about your experiences. I can correctly, fully, even vividly, represent them to myself. But my being able to represent to myself the feel of your experience is, in a way, too much of a good thing. It leaves me with one perception of the feel of my own experience and a second perception of the feel of yours. There is still a gap. How do I get the *two* experiences on to *one* scale ?

Hare has provided a more systematic statement on this point:

> The problem of ordinality versus cardinality is more difficult. On the face of it it looks as if our method does not require us to be able to measure utilities in constant units or with a constant zerto point. For it is enough if I, who am making a moral decision by critical thinking, can say 'Jones prefers outcome J_1 to outcome J_2 more than Smith prefers outcome S_2 to outcome S_1'. We do not have to be able to say how much more. This is because in our method of critical thinking we are not summing utilities, but, from an impartial point of view, which treats Jones' and Smith's equal preferences as of equal weight, forming our *own* preferences between the outcomes. (Hare 1981, 123)

But notice that, in this quoation, Hare is already assuming two things: First, *we already know* whether or not a preference of Jones' is equal (in strength) to a preference of Smith's; Second, we form our own preference from an *impartial point of view.* If we wish to provide a criterion of interpersonal comparison of preferences, the First begs the question. And the Second seems to be nothing but a restatement of

Sidgwick's Principle of Rational Benevolence, although it is formulated in terms of preferences. Recall that, according to Sidgwick, if we treat one's own good impartially through time, it is the Principle of Prudence (Rational Self-Love), and if we treat different persons's good impartially, that is the Principle of Benevolence (and let me add that Sidgwick's notions of good and pleasure already incorporate preferences; see Sidgwick 1907, 110-1, 127*).

---

*Sidgwick 1907, 110-1:

It would seem then, that if we interpret the notion 'good' in relation to 'desire', we must identify it not with the actually *desired*, but rather with the *desirable*: —— meaning by 'desirable' not necessarily 'what *ought* to be desired' but what would be desired, with strength proportioned to the degree of desirability, if it were judged attainable by voluntary action, supposing the desirer to possess a perfect forecast, emotional as well as intellectual, of the state of attainment or fruition.

*Sidgwick 1907, 127:

but, for my own part, when I reflect on the notion of pleasure, —— using the term in the comprehensive sense which I have adopted, to include the most refined and subtle intellectual and emotional gratifications, no less than the coarser and more definite sensual enjoyments, —— the only common quality that I can find in the feelings so designated seems to be that relation to desire and volition expressed by the general term "desirable", in the sense previously explained. I propose therefore to define Pleasure —— when we are considering its "strict value" for purposes of

quantitative comparison —— as a feeling which, when experienced by intelligent beings, is at least implicitly apprehended as desirable or —— in cases of comparison —— preferable.

As Hare himself mentions after the preceding quotation, cardinal (quantitative) utility is obtained if we accumulate enough (infinite, though) preferences. And, further, let us notice that Hare spells out the content of this impartiality, in his reply to Griffin in Seanor and Fotion (1988), as follows:

> They [informed preferences] are formed 'from scratch . . . from an understanding of the objects before us'. With this I agree; but (...) 'understanding' means understanding of their nature, not of their objective or generally accepted value; and though the judger judges from his own particular point of view, the exclusion of appeal to his own antecedent preferences or values means that any judge who truly represented to himself the situations-cum-preferences would form the same order of preference. This order of preference is thus objective in the sense that all rational informed judges, judging from a universal point of view which excludes their own other preferences, will share it. (Hare 1988, 238)

Notice how close Hare comes to Sidgwick in the last part of this quoation. In a word, if we have full information and form a preference, excluding our own personal preferences, that's a preference from an impartial point of view, and *it is the same for everyone* with the same conditions. I just wonder *where* Hare has established that this in fact holds! I cannot agree with him that 'the exclusion of appeal to his own antecedent preferences or values means' that everyone has the same

ordering; this is not a proof, but merely an assumption.

It is of course permissible to *call* this impartiality (with respect to preferences) another kind of universalizability. But as I have already pointed out by analogy to the measurement of length, this new kind of universalizability is concerned with *new concepts*, i.e. the *strength* of *preference*, which enables us to compare different people's preferences (assuming, for the sake of argument, that the strength is somehow ascertained), clearly distinct from the Principle of Justice or from 'the same facts, the same Ought' sort of principle. In forming our preferences (i.e. preferences all told) in critical thinking, it is logically possible to assign a different weight to my own or to other's preferences (that's why Hare has to mention 'an impartial point of view'); even if I represent other's preferences by my own (which is the job of the Conditional Reflection Principle), I can still distinguish them from my original preferences (noticethat Hare is assuming Archangelic knowledge in critical thinking) so that it is up to me with what weight I should treat them. Harsanyi is quite clear about this when he discusses the 'conversion rations' (between different utility functions) when we form a social-welfare function from individual utility functions (Harsanyi 1977, 57). Even if each individual has a cardinal utility function, one's *unit of utility* may be different from another's, and that's why we need coversion rations for a social-welfare function. Notice that such ratios amount to weights of preferences. Thus, whether we call the impartiality in question 'universalizability' or 'benevolence' or whatever, the fact remains that it is non-tautological and needs substantive justification; you may attribute it to the concept of morality, but then the question 'why should we be moral ?' becomes heavier and nothing is improved.

In short, even if we allow Hare to assume the 'correct

representation' of others' preferences, including the strength of these preferences, he has to face the following dilemma: *Either* (1) the strength of each individual's preference does not have a common scale or unit, *or* (2) it has; but if (1), Hare cannot meaningfully talk about an impartial treatment of everyone's preferences, and if (2), he still has to assume a unique way to assign a weight to each individual, i.e. the impartial weight, which cannot be justified on logical grounds alone. Further, if in case (2) Hare tries to derive the impartial weight from *rationality*, Hare is no different from Sidgwick.

8. Conclusion

Thus, I have to conclude that Hare, even for his simplest model case, has to assume analogues of *all* the three principles of Sidgwick. First of all, Hare's strong universalizability is non-tautologous and as strong as Sidgwick's Principle of Justice. Secondly, Hare's requirement of impartiality through time with respect to one's own preferences is nearly as strong as Sidgwick's Principle of Prudence; I say 'nearly' because Hare is not explicitly committed to a quantitative notion of good or utility, but the requirement of *equal weight* to the strength of each *preference* is quite distinct from the universalizability of 'ought' or any other evaluative words. This holds even if we pass over the problem of representing preferences at other moments now. Thirdly and finally, the requirement of equal weight to the strength of one's and other's preferences in critical thinking is also nearly as strong as Sidgwick's Principle of Benevolence, with the analogous proviso, applying to the representation of other's preferences within oneself. Briefly, it is one thing to represent other's preferences in oneself (by the Conditional

Reflection Principle, which I did not question, for the sake of argument, in this paper), and it is quite another thing to treat them equally or impartially; the latter amounts to an essential evaluative priciple for Sidgwick's utilitarianism.

Hare's use of the word 'universalizability' has tended to conceal these problems under the name of 'logic'. Reading Sidgwick anew, I came to this conclusion. However, since Hare's method of justification of moral judgment still seems clearer than Sidgwick's intuitionist way (though he adds some elaboration in Sidgwick 1879), I do not mean to abandon Hare's method altogether; only we need to notice where we are assuming substantive principles.

Note: Although my interpretation of Sidgwick's three principles and my criticism of Hare's theory based on it are original (it occurred to me during the summer of 1997, and was communicated briefly in the Sidgwick Mailing List, early September), Mariko Okuno has already written, on my suggestion, another version utilizing the same ideas in her Ph.D. Thesis; she named the interpretation 'Uchii-Okuno Interpretation' giving the main credit to me. Although I am happy with this name, and agree with most of what she says, I did not use that name in this paper. However, I wish to acknowledge that I had the benefit of examining her thesis and learning a great deal from her analysis of Sidgwick's view before I prepare this paper; and I wish to thank her for helpful comments on this paper. For her version, see Okuno 1998b, 7.2 and 9.3. She has also argued for the significance of Sidgwick's hedonism; see Okuno 1998a and 1998b, 10.1-10.4.

Bibliography

Gibbard, A. (1988) 'Hare's Analysis of "Ought" and its Implications', in Seanor and Fotion 1988.

Griffin, James (1988) 'Well-being and its Interpersonal Comparability', in Seanor

and Fotion 1988.

Hare, R. M. (1952) *The Language of Morals.* Clarendon Press, 1952.

——— (1955) 'Universalizability', *Proceedings of the Aristotelian Society* 55 (1954-55).

——— (1963) *Freedom and Reason.* Clarendon Press, 1963.

——— (1981) *Moral Thinking: its levels, method, and point.* Clarendon Press, 1981.

——— (1988) 'Comments', in Seanor and Fotion 1988.

Harsanyi, J. C. (1976) *Essays on Ethics, Social Behavior, and Scientific Explanation.* Reidel, 1976.

——— (1977) *Rational Behavior and Bargaining Equilibrium in Games and Social Situations.* Cambridge University Press, 1977.

——— (1982) 'Morality and the Theory of Rational Behaviour', in Sen and Williams 1982.

Okuno, Mariko (1998a) 'Rethinking Sidgwick's Hedonism' [in Japanese], *Arche 6,* 1998.

Okuno, Mariko (1998b) *Sidgwick and the Contemporary Utilitarianism* [in Japanese], Doctorate Thesis, Graduate School of Letters, Kyoto University, 1998.

Reichenbach, H. (1958) *The Philosophy of Space and Time.* Dover, 1958.

Schneewind, J. B. (1977) *Sidgwick's Ethics and Victorian Moral Philosophy.* ClarendonPress, 1977.

Sen, Amartya and Williams, B., eds. (1982) *Utilitarianism and Beyond.* Cambridge University Press, 1982.

Seanor, Douglas and Fotion, N., eds. (1988) *Hare and Critics.* Clarendon Press, 1988.

Shionoya, Yuichi (1984) *The Structure of the Idea of Value* [in Japanese]. Toyo-keizai-shinpo, 1984.

Sidgwick, Henry (1879) 'The Establishment of Ethical First Principles', *Mind* 4, 1879.

Sidgwick, Henry (1907) *The Methods of Ethics,* 7th ed. Macmillan, 1907.

Singer, Peter (1974) 'Sidgwick and Reflective Equilibrium', *The Monist* 58, 1974.

Uchii, Soshichi (1974) 'On the Universalizability of Moral Judgements' [in Japanese], *The Zinbun Gakuho* 38, Institute for Humanistic Studies, Kyoto University, 1974.

——— (1988) *The Law of Freedom, the Logic of Interest* [in Japanese]. Minerva,

1988.

———(1994) 'Expository Essay and Epilogue' [in Japanese], in Japanese translation of Hare 1981 by Uchii, S. and Yamauchi, T., Keiso, 1994.